editorial Journal of proteome research

Using Chemometrics and Statistics To Improve Proteomics Biomarker Discovery

The ability to measure large numbers of proteins and protein fragments—proteomics—has the potential to speed discovery of biomarkers that are useful for studying pathologic mechanisms, for disease diagnosis, and for prognostication after disease onset. However, several recent articles have called into question the reproducibility and relevance of reported proteomics biomarkers (*1–4*) and pointed to the absence of positive validation studies in the literature. Federal agencies, including the National Cancer Institute (NCI), recognize the potential for proteomics technologies in disease diagnosis and therapeutic monitoring; thus, they are taking steps during both the discovery and validation phases to improve the reliability of biomarkers. These steps include establishing programs to identify sources of experimental variation, to develop standards and protocols, and to assess experimental designs intended to improve interlaboratory and intralaboratory reproducibility (http://proteomics.cancer.gov). Participants in these initiatives will be encouraged to have statistical input at all stages of their experiments.

Numerous scientific advances are aided and enhanced by the participation of the statistics and chemometrics communities. Chemometrics, which the International Chemometrics Society defines as "the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods," can make an important contribution to proteomics research. Chemometrics facilitates the extraction of chemically relevant information by optimizing experiments; processing data; and calibrating, organizing, and performing quality-control checks on the analytical process. Chemometrics is cross-disciplinary, as it incorporates the chemical, mathematical, and computational sciences.

One area where statistics plays an important role, and one of the first problems addressed by chemometrics, is the evaluation of interlaboratory variation to ensure that novel instruments have adequate measurement accuracy and precision and that they can be recommended to a broader scientific community as research tools. Pioneers such as Jack Youden, a chemist by training; John Keenan Taylor, also a chemist; and John Mandel, a chemist–statistician, established many of the key principles and experimental designs, such as Youden squares (*5–8*). The role of statisticians and chemometricians in formulating and implementing a proper experimental design is fundamental to ensuring the validity and credibility of interlaboratory comparisons and the quality of the measurements.

A typical MS-based proteomics biomarker discovery project consists of study design, data acquisition, data mining, and characterization of selected biomarkers. Statisticians and chemometricians can help researchers identify attainable goals and ensure that proposed approaches can meet those objectives. Statisticians can provide input on study-design issues, such as sample selection and the number of samples to be used. They can also help identify and control the factors that cause variation in the experimental outcomes, such as the type of mass spectrometer, the participating labs, sample handling, storage conditions, and population variation. Identification of systematic errors is especially important as it will lead to the design of new, more informative studies.

Chemometricians have addressed many of the issues regarding data acquisition from MS instruments, including baseline subtraction, calibration, normalization, alignment, and peak detection. For example, a substantial chemometrics literature exists on the identification, location, and fitting of peaks (e.g., *9, 10*). Once data have been properly standardized and prepared for analysis, statisticians can play an important role in data mining to attempt to reliably evaluate associations of proteomics patterns with biologic processes or disease. This involves the preselection of promising peaks, statistical model building, and comprehensive evaluation of sources of systematic and random error; in addition, it frequently requires resampling techniques to obtain a realistic assessment of exploratory findings. Statistical analyses to address key hypotheses should be specified in the study protocol; however, exploratory approaches may allow one to formulate important new hypotheses and may lead to further experiments and improvements in experimental approach.

Systems that integrate computational sciences, data processing, and statistical tools can facilitate biomarker discovery. For example, in a collaborative effort between the Fred Hutchinson Cancer Research Center and NCI, a comprehensive, web-based software platform called the Computational Proteomics Analysis System (CPAS) was developed for exploratory data analysis (*11, 12*). CPAS provides researchers with open-source tools for organizing, managing, processing, and interpreting the vast amounts of data generated by proteomics experiments.

In addition to participating in the design and interpretation of proteomics studies, statisticians and chemometricians can assist in the evaluation of papers in peer review. They can make sure that the study design and reported data justify the conclusions drawn, and they may be able to suggest more appropriate analyses of the available data or identify the need for additional data. In cases where a paper cannot be sufficiently corrected to support important and valid conclusions, the statistician can alert the editor. Finally, statisticians and chemometricians can help make a checklist for authors to consider before submitting a paper for publication. For example, such a checklist may include guidance on how to describe the experimental design and the analytic methods.

Involving statisticians and chemometricians in the design, implementation, and interpretation of proteomics studies, as well as in the editorial process, has the potential to improve the clarity and incisiveness of work in this field and to facilitate scientific advances. Rigorous application of sound statistical and chemometric principles will benefit the overall scientific community by improving protein biomarker discovery and validation.

Clifford H. Spiegelman
Texas A&M University
Ruth Pfeiffer and Mitchell Gail
National Cancer Institute

References

(1) Coombes, K. R.; et al. Serum proteomics profiling—a young technology begins to mature. *Nat. Biotechnol.* **2005**, *23*, 291–292.

(2) Ransohoff, D. F. Lessons from controversy: ovarian cancer screening and serum proteomics. *J. Natl. Cancer Inst.* **2005**, *97*, 315–319.

(3) Diamandis, E. P. Proteomic patterns to identify ovarian cancer: 3 years on. *Expert Rev. Mol. Diagn.* **2004**, *4*, 575–577.

(4) Baggerly, K. A.; et al. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J. Natl. Cancer Inst.* **2005**, *97*, 307–309.

(5) Youden, W. J.; Connor, W. S. New experimental designs for paired observations. *J. Res. Natl. Bur. Stand.* **1954**, *53* (3), 191–196.

(6) Youden, W. J. Physical measurements and experimental design. In *Colloques Internationaux du Centre National de la Recherche Scientifique, No. 100, Le Plan d'Experiences*, 1961; pp 115–128.

(7) Taylor, J. K.; Cihon, C. *Statistical Techniques for Data Analysis*, 2nd ed.; Chapman & Hall/CRC Press: Boca Raton, FL, 2004.

(8) Paule, R. C.; Mandel, J. Consensus values and weighting factors. *J. Res. Natl. Bur. Stand.* **1982**, *87* (5), 377–385.

(9) Phillips, G. W.; Marlow, K. W. Automatic analysis of gamma-ray spectra from germanium detectors. *Nucl. Instrum. Methods* 1976, *137* (3), 525–536.

(10) De Braekeleer, K.; Torres-Lapasío, J. R.; Massart, D. L. Improved purity assessment of high-performance liquid chromatography diode array detection data for overcoming the presence of the non-linearity artefact. *Chemom. Intell. Lab. Syst.* **2000**, *52*, 45–59.

(11) Rauch, A.; et al. Computational Proteomics Analysis System (CPAS): An extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **2006**, *5*, 112–121.

(12) Cottingham, K. CPAS: A proteomics data management system for the masses. *J. Proteome Res.* **2006**, *5*, 14.